

Towards a Zero-to-Multi-Genre Classification Scheme

Marina Santini¹

¹University of Brighton (UK)

Abstract

In this presentation, I will argue that automatic methods for genre identification on the web should be more flexible and informative than merely assigning a single genre label to a web page. As the genre system on the web is a complex reality, and web pages are much more unpredictable and individualized than paper documents, I propose an inferential model that permits a classification that can accommodate genres that are not entirely standardized, and is more respectful of the actual nature of a web page, which is mixed, rarely corresponding to an ideal type and often showing a mixture of genres or no genre at all. A proper evaluation of such a model remains an open issue.

Résumé

Étant donné que le système des genres sur le web est complexe et que les pages web sont plus imprévisibles et individualisées que les documents traditionnels, nous proposons un modèle déductif-inductif qui permet une classification qui peut s'accommoder des genres non complètement standardisés. Il est aussi plus respectueux à l'égard de la vraie nature de la page web, qui est en fait mixte et ne correspond presque jamais à un type idéal ou à un prototype précis, mais présente plutôt un mélange de genres, ou pas de genre du tout. L'évaluation de ce modèle reste un problème à résoudre.

Mots-clés : genre, typologies textuelles, pages web, modèle déductif-inductif, identification automatique, théorème de Bayes.

1. Introduction

It has been often pointed out that genre is a difficult concept to pin down, and automatic classification by genre is difficult because genres are heterogeneous categories. Although a considerable amount of research has already been done in automatic genre classification, most previous work has considered genres as mutually exclusive categories, disregarding the fact that many documents, and in particular many web pages, cannot be fitted into a single genre. This approach has been taken for the sake of practicality, but proves inadequate when dealing with complex documents, like web pages.

In particular, two factors affect genre identification on the web: (1) the complexity of web pages, and (2) the fluidity and the fast-paced evolution of the web.

First, genres on the web are instantiated in web pages which, from a physical, linguistic and textual point of view, can be considered documents of a new type, much more unpredictable and individualised than documents on paper. Web pages show a visual organization of the space, where several communicative purposes are included at the same time. In fact, the space in a web page is often divided into different sections, organized by snippets of text scattered around the main body of the document, like navigational buttons, menus, ads, and search boxes. In brief, in a web page, not all the elements necessarily belong together, but they all contribute to form a whole.

Second, the web is a recent communication medium, invented only in 1989. Being so recent, it is still fluid and unstable. In addition, it is evolving at a fast pace owing to the continuous introduction of new web technologies. In such an environment, the modification of existing genres or the creation of new ones, which better satisfy the communication needs brought about by the new conditions, are common phenomena. One effect of the fast evolution is the presence of emerging genres.

Emerging genres represent a transitional phase in genre evolution. They can be considered as hypothesised genres. They are still in an embryonic form, and it is not yet clear whether they will ever coalesce into a new communication object. The concept of emerging genres is useful because it accounts for unclassified or unclassifiable web pages, and has some practical benefits. For example, it can help organize the large amount of unclassifiable pages populating the web.

In summary, web pages are noisy documents and the web is a noisy environment. This noise results in classification hurdles, because many web pages have often more than one genre or do not have any. If we aim at devising an automatic classification system capable of facing the *open web*, and not only a closed-world environment, we should take these factors into consideration. The open web accounts for a situation where the population is unknown, where web pages might be evolving, hybrid, individualised, or not following any genre convention.

For these reasons, I propose a more flexible genre classification scheme based on an articulated genre analysis, capable of assigning zero, one or multiple genre labels (zero-to-multi genre classification), by leveraging on the linguistic features returned by NLP tools.

This scheme relies on the observation – put forward by Werlich (1976) and other textlinguists, and confirmed by genre-specific analyses – that genres are characterized by recurrent or predominant text types. Text types are linguistic devices expressing the purpose of communication. Textlinguists and some genre analysts (for example, Paltridge, 1996) use text types to analyse genres. However, textlinguistics and genre analysis mostly lack corpus-based evidence and are not computational. For this reason, I propose exploiting the observations about the relation between text types and genre, and test them using a corpus-based approach, enlarging the range of automatically-extractable features, and implementing statistical and computational methods that incorporate findings from neighbouring fields, such as automatic genre classification, textlinguistics, corpus-linguistics, and genre analysis. Biber has already implemented a corpus-based approach to automatic text type identification (Biber, 1988), but his methodology focuses on text type detection across genres rather than on automatic genre identification within the individual document.

In the model that implements the zero-to-multi genre classification I make a clear-cut separation between the concepts of text types and genre. Text types are rhetorical/discourse patterns that indicate the purpose of communication, like *NARRATION*, *INSTRUCTION*, and *ARGUMENTATION*. When the purpose is to narrate, the *NARRATION* text type is used; when the purpose is to instruct, the *INSTRUCTION* text type is used; and so on. Normally, a text contains several purposes and consequently includes several text types. Text types tend to be universal, and cut across time, cultures and societies. Genres are, on the other hand, culturally defined and socially acknowledged text categories, like *EDITORIALS*, *TUTORIALS*, *HOME PAGES* or *BLOGS*. It is important to stress that genres are linked to a particular historical context, and are in constant evolution. The interaction of these two concepts allows for more flexibility because (i) a classification in terms of text types remains possible even if a web page belongs to an emerging genre and cannot be assigned to any existing genre (zero-genre assignment),

and because (ii) a web page can be labelled with more than one genre (multi-genre assignment) when some genres share the same text types, as in the case of EDITORIALS and SERMONS, which are both *ARGUMENTATIVE*.

2. Methodology: The Inferential Model

The model that I will present in the following subsections is based on inference and not on machine learning. For this reason, I refer to this model as *the inferential model*. The methodology relies on the following elements : (i) a corpus of web pages that approximate the web, (ii) facets (i.e. linguistically-motivated features), and (iii) a modified version of Bayes' theorem, i.e. the odds-likelihood or subjective Bayesian method (Duda et al., 1981).

2.1. The Web Corpus

Since the web is in constant flux, it is almost impossible to compile a representative corpus/sample. The solution that I suggest for this model is to approximate one of the possible compositions of a random slice of the web statistically supported by reliable standard error measures. The distribution of the web corpus must be *approximately normal*. This fact is very important because the inferential model is based on z-scores, and z-scores assume a normal distribution of the sample. My web corpus contains 2,480 web pages. Out of 2,480 web pages, one part (1,000 web pages from the SPIRIT collection, i.e. around 40%) is non-annotated by genre, while another part is annotated by genre (1,480 web pages, i.e. 60%). The evaluation will be carried out on 1,400 web pages belonging to seven web genres, namely : BLOGS, ESHOPS, FAQs, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES, and SEARCH PAGES.

2.2. The Facets

I use the term 'facet' to indicate an 'aspect' in the communicative context that is reflected in the use of the language. The denomination of 'facet' emphasizes the fact that linguistic and textual traits represent an aspect of communication that can be *interpreted* for deriving text types or genres. For this reason, I define my facets as *functionally-interpreted features*. My facets are macro-features. That is, a facet may contain several micro-features. For example, the *first person facet* includes all first person pronouns, i.e. singular, plural, possessive and reflexive. All in all, I use 100 facets. A comprehensive description of facets can be found in Santini (2005).

2.3. Odds-Likelihood

Like the standard Bayesian version, the odds-likelihood method is based on probabilities. Odds and probabilities contain exactly the same information and are interconvertible. But odds are not limited to the range 0-1, like probabilities. Odds is a number (without any limitation) that tells us how much more likely one hypothesis is than the other. The main difference between the regular Bayes models and the subjective one is that in the latter attributes are NOT considered to be equally important, but are, instead, weighted according to their probability value. Therefore, in the odds-likelihood version of Bayes' theorem much of the effort is devoted to *weighing* the contributions of different pieces of evidence in establishing the match with a hypothesis. These weights are confidence measures: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist (larger value means greater sufficiency), while LN is used when evidence is known NOT to exist (a smaller value means greater necessity). $LS(E|H)$ expresses how much the prior odds,

$O(H)$, in presence of a clear evidence of E has to be multiplied in order to get the posterior odds, $O(H|E)$. LS is typically a number > 1 . $LN(E|H)$ expresses how much the prior odds, $O(H)$, in presence of a clear evidence against E has to be multiplied in order to get the posterior odds, $O(H|E)$. LN is typically a number < 1 . Usually $LS*LN=1$. In this implementation of the model, LS was set to 1.25 and LN was set to 0.8 on the basis of previous experience and empirical adjustments.

2.4. Steps

In the model, the individual web page as a whole is taken as unit of analysis, without removing any textual component. The model includes the following steps:

- Automatic extraction of functionally-motivated features – the facets – from a corpus representing the web.
- Inference of four text types – *DESCRIPTIVE_NARRATIVE*, *EXPLICATORY_INFORMATIONAL*, *ARGUMENTATIVE_PERSUASIVE* and *INSTRUCTIONAL* – from linguistic facets using odds-likelihood. The combination of facets in text types is based on previous studies (mainly Werlich, 1976).

Inferred text types are associated to a probability value. For example, a web page can have 0.9 probabilities of being argumentative, 0.7 of being instructional and so on. Probabilities are interpreted in terms of *degree* or *gradation*. For example, a web page with a probability of 0.9 of being argumentative shows a very high degree, or gradation, of *ARGUMENTATION*. The different gradations are independent from each other. In other words, the different probability values accounting for the four text types in a web page do not sum up to 1.0, but they simply indicate the *gradation*, and not the proportion, of a certain text type.

- Gradations/probabilities are then ranked in descending order (the highest probability gets the first position).

After the ranking, the following hypothesis is tested: the combination of two predominant text types, i.e. the top-ranked text types, plus a combination of additional traits, e.g. layout and functionality, is sufficient to derive seven web genres – BLOGS, ESHOPS, FAQs, FRONT PAGES, LISTINGS, PERSONAL HOME PAGES, and SEARCH PAGES. This hypothesis is tested with *if-then* rules. The combination of text types with other traits is mostly inspired to previous analyses of web genres. There is no special reason for combining only two predominant text types instead of three or more. The basic assumption is that web pages are mixed. Obviously, web pages may contain many other text types, not only the four broad text types included in the model. This text typology is a starting point and can be enlarged in future.

2.5. Evaluation of Results on Single-Genre Classification

The inferential model cannot be fully evaluated at this stage of genre research. Here, I evaluate the results on single-genre classification, deferring to future work a more comprehensive assessment.

I compared the outcome of the model on a single genre with the SVM and Naive Bayes classifiers from the Weka machine learning workbench. SVM and Naive Bayes classification models were built using only the seven web genres and not the entire web corpus. This means that I used 200 web pages per web genre, amounting to a total of 1,400 web pages. The stratified cross-validated accuracy returned by these classifiers for seed 1 is about 89% for SVM, and about 67% for Naïve Bayes. The accuracy achieved by the inferential model is

about 86%. An accuracy of 86% is a good achievement for a first implementation, especially if we consider that the standard Naive Bayes classifier returns an accuracy of about 67%. Although SVM achieves an accuracy of about 89%, i.e. about +3% than the inferential model, it is worth stressing that both Naive Bayes and SVM standard classifiers were run on 1,400 web pages. The inferential model, on the other hand, is built on a corpus of 2,480 web pages. This means that the inferential model can stand a certain level of noise, represented by web pages that have not been classified by genre. In order to see to what extent a supervised classifier can handle the unclassified web pages that might be found on the web, I built an 'unknown' class, i.e. a class that collects all the genres not included in the 7-web-genre palette. This means that I built an SVM classifier with the 2,480 web pages of the web corpus and labelled as DONTKNOW 1,080 web pages. This SVM model built with eight classes returns an accuracy of about 76%. In this respect, the inferential model appears to be much more accurate and stable than the supervised model, because it returns an accuracy of about 86% on 2,480 web pages, i.e. about +10% more than the model built with SVM on eight classes.

3. Conclusion

Although I argue that a zero-to-multi-genre classification scheme is more suitable for the current situation of genres on the web, where it is often difficult to fit a web page into a single genre, many questions remain unanswered. For example:

- Is possible to build an agreed upon genre-annotated reference corpus for evaluation purposes, currently missing? What criteria should be followed when annotating a web page by genre? Should multiple genres be rated? What criteria should be followed for evaluating emerging genres?
- Is it possible to create of a hospitable genre classification system, i.e. a system that is capable of hosting new genres, or updating existing ones, without upsetting the whole framework?
- Is it possible to integrate genres in a large-scale search engine?

These and many other issues will be addressed by future research.

Références

- Biber D. (1988). *Variations across speech and writing*. Cambridge University Press, Cambridge.
- Duda R., Hart P. and Nilsson N. (1981). Subjective Methods for Rule-Based Inference System. In Weber B. and Nilsson N. (eds.), *Readings in Artificial Intelligence*, Tioga Publishing Company.
- Paltridge B. (1996). Genre, text type, and the language learning classroom. *ELT Journal*, vol. 50(3): 237-243.
- Santini M. (2005). Linguistic Facets for Genre and Text Type Identification: A Description of Linguistically-Motivated Features. Technical Report ITRI-05-02. University of Brighton, Brighton (UK).
- Werlich E. (1976). *A Text Grammar of English*. Quelle & Meyer.